JOHANNES KEPLER
UNIVERSITY LINZ

**Tobias Höller**
Institute of
Networks and Security

@ tobias.hoeller@ins.jku.at
https://www.digidow.eu/

February 27, 2021

# Collecting statistical information on v3 onion services

Technical Report

Christian Doppler Laboratory for
Private Digital Authentication in the Physical World

This work has been carried out within the scope of Digidow, the Christian Doppler Laboratory for Private Digital Authentication in the Physical World, funded by the Christian Doppler Forschungsgesellschaft, 3 Banken IT GmbH, Kepler Universitätsklinikum GmbH, NXP Semiconductors Austria GmbH, and Österreichische Staatsdruckerei GmbH.

INSTITUTE
OF NETWORKS
AND SECURITY

DIGiDOW

**JOHANNES KEPLER**
**UNIVERSITY LINZ**
Altenberger Straße 69
4040 Linz, Austria
jku.at

# Contents

## 1. Introduction

Monitoring the activities of onion services by deploying multiple HSDir nodes has been done repeatedly in the past. With v3 onion services, Tor mitigated such attacks by blinding the public keys of onion services before uploading them. This effectively prevents the collection of onion addresses, but it does not prevent the collection blinded public key uploads and downloads, which provide statistical insight into how onion services are being used. Additionally, it is possible to identify and link blinded keys derived from well-known onion services, providing a solid estimate on how often they are accessed. This report presents our setup to collect statistically significant information on v3 onion service usage without compromising the privacy of Tor users.

## 2. Experiment Setup

We deployed a series of about 50 tor relays which meet the requirements for obtaining the HSDir flag (stable + uplink of at least 100kbit/s). All relays and the database they collect their information in, run within the infrastructure and IP address range of our university and remain under our control. For every blinded onion key that is uploaded to one of our nodes, we store the blinded key along with the node it was uploaded to and the hour it was uploaded in. We count both the number of repeated upload requests sent to the same node (as they are relevant to our security considerations) and the total number of downloads across all nodes. In order to provide transparency, all relays carry the correct family information and be labeled with a link to a website of our research project (If you have any suggestions/examples on how to best label our nodes, we would welcome your feedback). In parallel a list of known V3 descriptors will be obtained to calculate their blinded version for every time period where our HSDir nodes collected data.

## 3. Experiment Goals

- Find out adaption rate of V3 onion services (especially now that V2 is about to be deprecated).
- Obtain statistically relevant data on how frequently known V3 onion services are being used.
- Analyze how much of onion service usage is caused by well known onion services.
- Previous research has shown that some botnets use onion services to hide their Command&Control servers. This project could (once the address has been found) provide an estimate on how large such botnets are. Because the blinded descriptors are saved, this could even be done retroactively.
- Enable a more reasonable discussion of the “darknet”. Many people have completely wrong ideas about it (see that iceberg analogy that refuses to die).

# 4. Safety aspects

Protecting Tor users is a critical requirement when collecting information from a network that other users depend on to protect their privacy.

## 4.1 Ethical considerations

There are currently about 4000 Tor relays acting as HSDirs. While the Tor project is trying to identify HS directories which harvest onion addresses, this is typically done by uploading unpublished descriptors and checking if they are being accessed. As long as we do not probe V2 onion descriptors and return correct information (no manipulation of the DHT that is used for onion services) our relays would not fall under the current definition of "bad relays"1 (as suggested we will discuss this with the bad-relays team before deploying our nodes). Even if our actions would be considered malicious, it would be very hard to identify our behavior based on the actions of our relays. The same is true for a state actor intent on monitoring the Tor network in secret. They would most likely not be foolish enough to reveal themselves by accessing every onion address they see. So it seems like a fair assumption that this data is already being collected somewhere. Doing this openly and publishing the results provides public information which can be used to improve the security of the Tor network and removes an advantage in knowledge from malicious entities.

## 4.2 Privacy considerations

The main data we are extracting from the Tor network is a list of blinded V3 onion keys and a count of how often these keys have been requested by clients. We do not store individual requests any longer than it takes to extract the blinded onion v3 key and its validity period. In order to ensure that the relative order to upload requests cannot be inferred from our data, upload requests will be sorted according to the blinded public key they provided before being inserted into our database. Regarding the publication of data, we would like to make our results (aggregated over a duration of at least 3 months) publicly available to enable other researchers to work with them. For blinded keys which we cannot link together (because we do not have the onion key) we will only aggregate how many of them there are and how many requests were made for them in total and on average (per service, per day, ...). We identified two potential risks in relation to publishing our aggregated results:

1. We might identify frequently used onion services, making them more likely to be attacked by malicious entities. In order to mitigate this issue, we will blind all onion addresses before our initial publication. In future work we plan to discuss and identify onion services for which usage numbers can be safely published.
2. The second risk we identified, is the possibility that barely used onion services might still be deanonymized based on our data. Therefore, we will treat onion services, which see little to no usage (since we have no data collected yet, we feel unable to specify a clear threshold at this point) like unidentified services and just publish their number as a whole without any further information about them. In future work we would like to investigate methods like Differential Privacy (like privcount1) to enable us to safely publish more of our data.

### 4.3 Reliability considerations

Since all our relays are publicly labeled, a malicious entity could easily notice if their onion service was being measured by us and send lots of fake queries to our HSDir relay to falsify data. While the anonymous nature of the Tor network does not allow us to prevent this kind of attack, we hope that outlier detection on our collected data will enable us to identify interference with our measurements.

### 4.4 Legal considerations

This research project is undertaken by the Johannes Kepler University in Austria and the legal defense is therefore adjusted to Austrian law. The operation of Tor relays (also with the HSDir flag) has been found legal, as we have no way to identify any illegal actions happening via these relays and cannot be found responsible for actions we do not know about. The data we collect might be considered personal data (if the onion service is operated by a private entity), but since we only publish data about onion services which have published their keys and receive external traffic, their right towards keeping their usage numbers private is most likely outweighed by our legitimate research interests. Lastly the data we collect and publish does not encourage or enable any criminal activities. The only risk to the data we collect is the possibility that our data is subpoenaed by Austria Law Enforcement. The University could only fight such a subpoena after it has been executed and the data has been collected. So to ensure that our data is secured against such attacks, we plan to encrypt collected data with an asymmetric key. The private key is password protected with a password known only to two employees on the project. As Austria has no law that forces individuals to reveal passwords to investigators, this should be sufficient to keep the collected data secure. We will document this clearly on our website, which will hopefully discourage law enforcement from even trying to get access to our data.

## 5. Final remarks

This report was shared with the Tor research safety board prior to the deployment of the first Tor relays. After several iterations they had no further objections with the proposal described in the previous sections. Ongoing information about the experiment can be found on the responsible website: https://www.digidow.eu/experiments/onion-stats/